Preventing Pump and Dump Schemes with Using Supervised Machine Learning Models: Relevance in Stock and Cryptocurrency Market

Signature Work Research Report

[instructions]

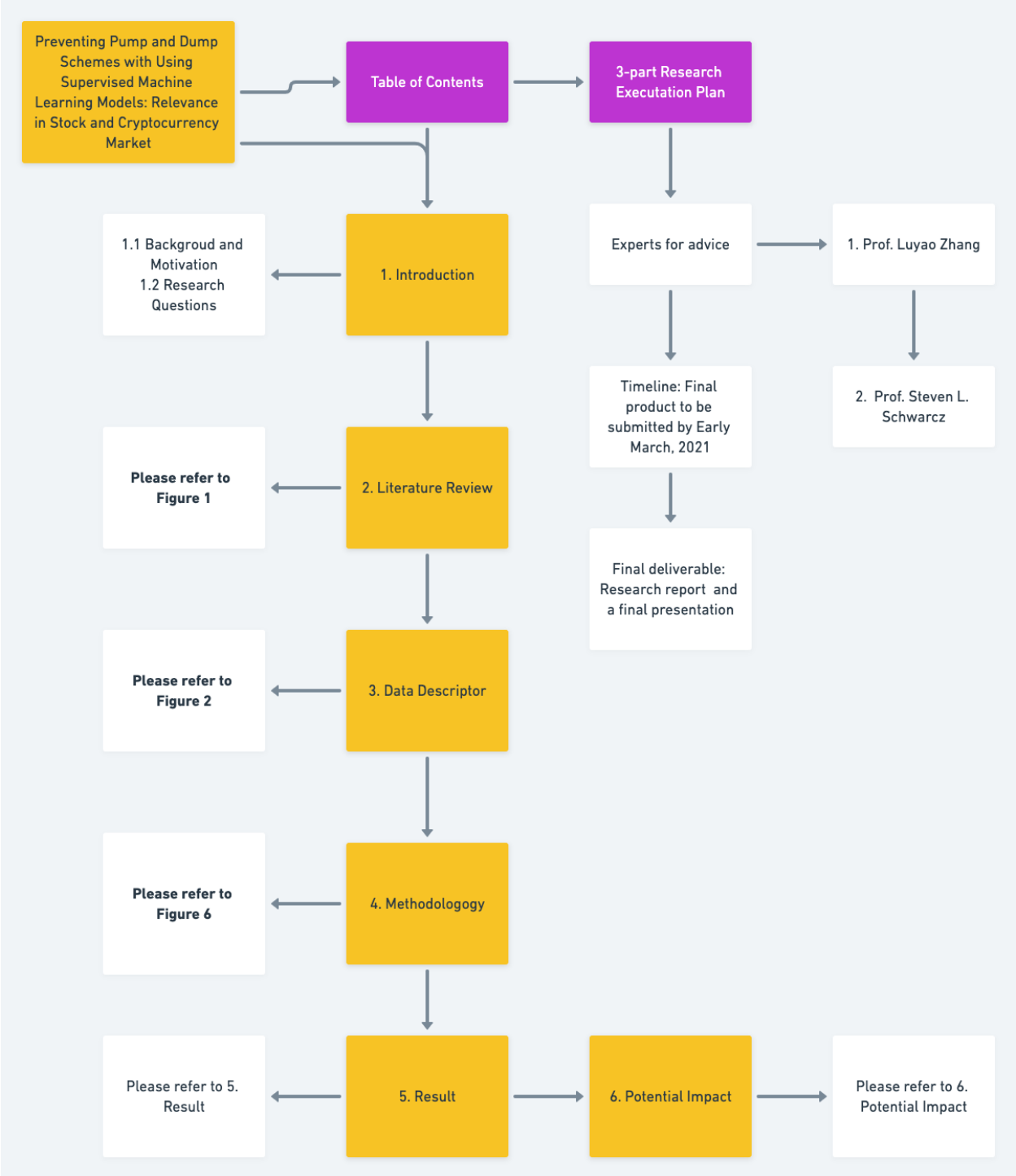
Author: Jennifer Lee, Economics, Class of 2022, Duke Kunshan University

Supervisor: Prof. Luyao Zhang, Duke Kunshan University

Keywords: pumps-and-dumps, artificial price manipulation, blockchain, cryptocurrency, stock market, compliance

Acknowledgments: The cryptocurrency pump-and-dump data used in this research is obtained from Jiahua Xu's openly accessible online repositories. One anonymous peer reviewer provided feedback to improve this paper. All the remaining errors remain to be mine.

NetID:jl873
Contact: jl873@duke.edu
Last Update: 12/13/201
Google Folder:
https://drive.google.com/drive/folders/1NccJUJEsrijT3HXJ0Oum-6_t0eef03bm?usp=sharing
Mybib Folder: https://www.mybib.com/j/ObservantFeignedWoodcock
Whimsical Folder: https://whimsical.com/sw-2dpaSfS47y13z9gbq9sjtg



Preventing Pump and Dump Schemes with Using Supervised Machine Learning Models: Relevance in Stock and Cryptocurrency Market Feature Photo (created with Whimsical)

Table of Contents

Abstract

Pump and dump schemes, where prices of assets are artificially manipulated by a few and dumped shortly after are a long-lasting problem on exchanges. This problem has been worsened in cryptocurrency exchanges as cryptocurrency remains largely unregulated as of yet. The current academic literature contains much research on cryptocurrency pumps-and-dump detections, utilizing various machine learning methods. However, there is not yet research on comprehensive machine learning model selection and each model's efficiency. There is also not much research on the best predictors of cryptocurrency pumps and dumps. Through our research, we will be able to identify the best-supervised machine learning model for the purpose of predicting pump-and-dump schemes in the cryptocurrency market. Our research insights would add to creating a more transparent cryptocurrency market for all investors. (SWCE Link)

Part I: Research Description

1. Introduction

1.1 Background and Motivation

In a span of just a few years, the market capitalization of cryptocurrency or crypto-tokens has increased from almost zero to more than $1 trillion (Harvey 2022). While there could be debates as to whether cryptocurrencies are a valid asset class or not, there is no doubt that cryptocurrencies are a part of our daily lives as even mom-and-pop investors are starting to trade cryptocurrency. Unlike other securities such as equities and fixed income that are formally recognized under the financial law, cryptocurrency is not subject to surveillance and monitoring by regulators as of yet. This lack of surveillance and monitoring by the financial authorities is

making cryptocurrencies subject to frauds and scams, financially damaging vulnerable mom-and-pop investors.

Classic securities scam that is not novel in the cryptocurrency market is a pump-and-dump scheme. Pump-and-dump schemes are artificial price manipulations through the spread of misinformation and were outlawed in the United States in the 1930s (Gandal et al. n.d., 6).

Pump-and-dump schemes have evolved in the cryptocurrency market with the advances in social media. As a result, the pump-and-dump organizers and participants now only require a shorter time frame, mostly no longer than a few minutes, to manipulate prices. This means that pump-and-dump organizers are given more time to arrange other similar pump events.

Pump-and-dump schemes work like illustrated below.

- 1. scheme organizers firstly collect interested members on social media channels such as Telegram or Discord (Gandal et al. n.d., 2).
- 2. When enough members are deemed to have been gathered, organizers set some pre-pump rules, including how members are supposed to buy fast and not to sell below their purchase price.
- 3. Members are encouraged to tout the advantages of their target pump coin on other social media channels.
- 4. After a pump is announced by the organizers, the coin price starts to rise and within a span of a few minutes, the price falls back to its pre-pump level.

Li et al. (2019) indicate that pump-and-dump schemes allow for a wealth transfer between insiders that were part of pump-and-dump groups and outsiders that were unaware of such scheduled price manipulation. Given the significant size of financial damage done to vulnerable outsiders on a daily basis, it is critical for researchers to devise a prevention detector tool for wider commercial use. Further developing on the dataset provided by Xu and Livshits

(2019), we research what is the best supervised machine learning model for the pump-and-dump prediction model and find out the best model's accuracy.

Furthermore, as pump-and-dump schemes find their roots in the stock market, we want to also compare and contrast if supervised machine learning models are effective in the stock market by referencing the literature on pump-and-dump prevention models in the stock market.

More research on the pump-and-dump predictors is needed for practical real-world usage by regulators and retail investors.

We foresee that regulators could utilize and apply our model to help diminish the pump and dump scheme ecosystem. Furthermore, any investor could use our model to check if a specific cryptocurrency trading pair is subject to artificial manipulation in the near future.

1.2 Research Questions

We propose the following research questions that we aim to answer in our research. In the previous literature, these questions have only been superficially covered without a profound depth. Recognizing the need for research on which supervised model might have the best accuracy in predicting pump-and-dump schemes, we would like to address this current gap in the relevant literature.

- **What are the features or predictors important in detecting pump and dump schemes in the cryptocurrency market?**
- **What is the best machine learning model for the purpose of predicting pump and dump schemes in the cryptocurrency market?**
- **How effective is the selected supervised machine learning model in detecting pump and dump schemes in the cryptocurrency market?**
- **How does the efficiency of supervised machine learning models compare and contrast in the cryptocurrency market and the stock market?**

2. Literature Review

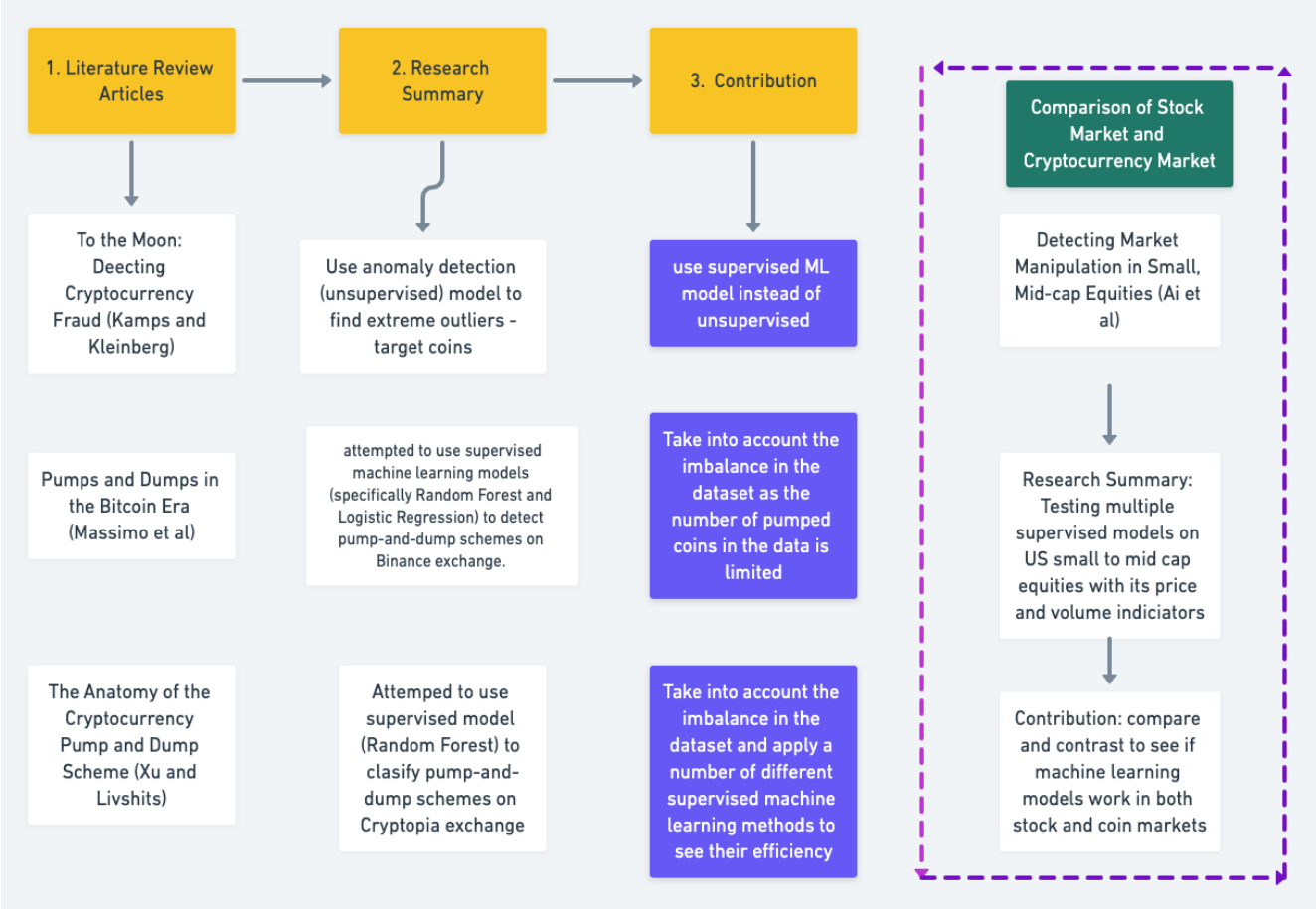


Figure 1. Literature Review Flowchart (Created with Whimsical)

In response to the urgent need for preventing pump-and-dump, there have been a few pump-and-dump detectors developed in the past. Kamps and Kleinberg (2018) pioneered the academic study on detecting pumps and dumps schemes by utilizing an anomaly detection model. The idea behind the anomaly detection model is that if the price and volume of a pump-and-dump target coin is above certain mathematical thresholds, then the coin is regarded as being pumped. Their methodology is paramount to finding extreme outliers in the data. One of the limitations of the study by Kamps and Kleinberg was that they used an unsupervised machine learning model based on queried market data. Therefore, their model was limited in confirming whether their detected pump-and-dump was actually indeed a result of such activity.

Massimo et al. (2020) attempted to use supervised machine learning models (specifically Random Forest and Logistic Regression) to detect pump-and-dump schemes on Binance exchange. The team led by Massimo generated their own research dataset that consists of confirmed Telegram pump-and-dump activities. They have found that Random Forest Classifiers perform better than a Logistic Regression through a series of 5 and 10 fold cross-validations. Also due to the fact that Massimo's team collected the dataset on their own, confirmed pump-and-dumps on Binance is only about 100.

Xu and Livshits have also developed a Random Forest Classifier model for predicting pump-and-dump with the data that they collected from various API sources, including CryptoCompare and PumpOlymp. A potential research area or limitation of the paper by Xu and Livshits is that they only use Random Forest Model without testing other supervised models first.

The common trend across the relevant pump-and-dump literature is that random forest supervised classifiers were the most widely used. A study by Shao (2021) also indicates that a random forest classifier has the best accuracy in detecting Dogecoin pump-and-dump schemes.

In contrast to using only 1 or 2 machine learning models (see Massimo et al and Xu and Livshits), it is important to compare the multiple supervised machine learning models (Random Forest classifier, K-Nearest Neighbors, Logistic Regression, Decision Tree classifier, AdaBoost classifier and etc) to find out which model actually has the best accuracy.

The study named "Detecting Market Manipulation in Small to Mid-cap Equities" explores how effective various supervised learning models, including Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression are in detecting pump-and-dump in American small to mid-cap equities (Ai et al. n.d.). As this study has researched similar research questions as our research, we want to compare and contrast the results to see how useful and

applicable machine learning would be in detecting pump-and-dump across cryptocurrency and stock market.

Overall, we propose a research on discovering the best supervised machine learning model with a cross-validation technique. Furthermore, in order to include in our dataset the confirmed pump-and-dump activities, unlike the study by Kamps and Kleinberg, we will only utilize supervised machine learning models, not unsupervised. In a pump-and-dump dataset, it is expected for the labeled pump to be much smaller than the labeled not pump. The study by Massimo et al. (2020) has only about 100 pumps to use in their model. In order to account for this weakness, we can balance the imbalance between the labeled pump and not pump by Synthetic Minority Oversampling Technique (SMOTE) technique. This will be explained further in the Methodology section.

3. Data Descriptor

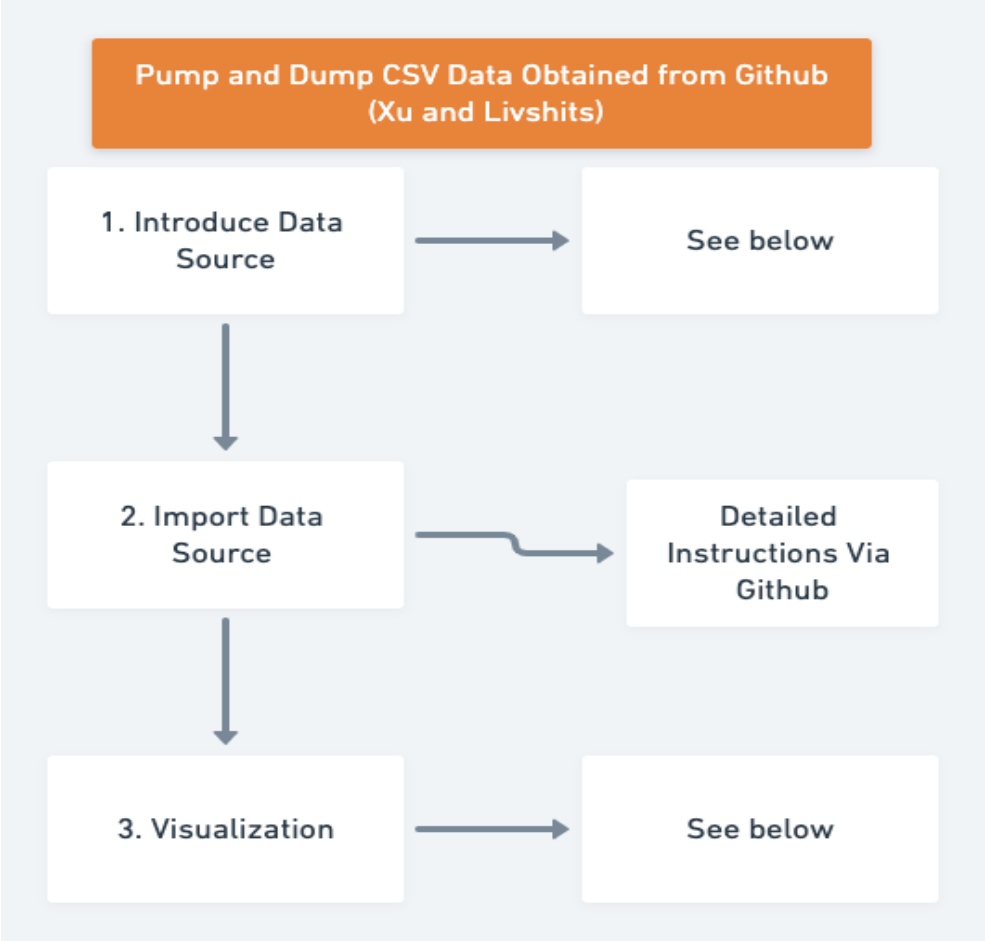


Figure 2. Data Descriptor Process Flowchart (Created with Whimsical)

3.1 Introducing Data Source

Usually, pump-and-dump schemes occur across multiple cryptocurrency exchanges, including Binance, Bittrex, Cryptopia, Yobit, and etc. Each exchange shows different characteristics in regards to their user base, their volume, and coins traded on the exchange. Naturally, as bigger exchanges such as Binance and Bittrex have more active users and coins that have relatively bigger market capitalization. In contrast, smaller exchanges such as Cryptopia and Yobit let their users trade less popular, small to mid-cap coins with lower liquidity. This

means that more pump-and-dump schemes are conducted on Cryptopia and Yobit as pump-and-dump organizers can artificially manipulate the price of coins with more ease.

Hence, the data source that we will be using in this research focuses on the pumps and dumps that occurred on Cryptopia. The authors that we have introduced in our research, Xu and Livshits, have shared their Cryptopia data publicly. Credits to Xu and Livshits, we will be utilizing their dataset in our research.

To explain how Xu and Livshits obtained their data initially, they first obtained pump-and-dump group data on the PumpOlymp website, which collects the information about pump-and-dump groups across different social messenger applications, including Telegram and Discord. After that, Xu and Livshits cross-validated the data obtained on PumpOlymp by manually checking the groups on Telegram. Per each pump-and-dump group found online, Xu and Livshits added various technical indicators of the pump target coin by querying the market data from CryptoCompare API (2019, 1615).

The data is about 70,239 rows long with non-pump and dump coins accounting for 70120 cases and pumped coins accounting for 119 cases.

Table 1 introduces the variables and features included in the dataset.

Variable Name	Description	Unit	Type	Frequency
pumpedtimes	How many times has the referenced coin been pumped on Cryptopia in the past	/	float	[0, +∞)
last_price	Price of the referenced coin one hour before the pump announcement	/	float	[0, +∞)

	• open price of the day			
views	How many times has the Telegram group been viewed	/	int	[0, +∞)
caps	Market capitalization of the referenced coin before the pump	/	float	[0, +∞)
volumefrom3h	The volume of the referenced coin traded before 3 hours to one hour before the pump announcement	/	float	[0, +∞)
volumeto3h	volume of the referenced coin traded (measured in BTC) from 3 hours to 1 hour before the pump	/	float	[0, +∞)
return3h	3-hour log return of the coin within the time window from 4 hours to 1 hour before the pump	/	float	[0, +∞)
returnvola3h	Volatility in the hourly log return of the referenced coin within the time window from 4 hours to 1 hour before the pump	/	float	[0, +∞)
volumetovola3h	The volatility in the hourly trading volume in BTC within the time window from 4 hours to 1 hour before the pump	/	float	[0, +∞)
pumped	The status of whether the	/	boolean	(0,1)

	referenced coin has been pumped or not 0- not pumped 1-pumped			

Table 1. Data Dictionary of the Dataset

3.2 Import Data (Data Availability)

For the data source and the github code, please refer to here.

3.3 Data Visualization

Figure 3 represents a side-by-side plot of the 3-hour return before the pump announcement with the pumped coin's 3-hour return on the left side and non-pumped coin's return on the right. An insight from Figure 3 is that even before a formal pump announcement on social media, pump target coins' volatility is greater than non-pump coins.

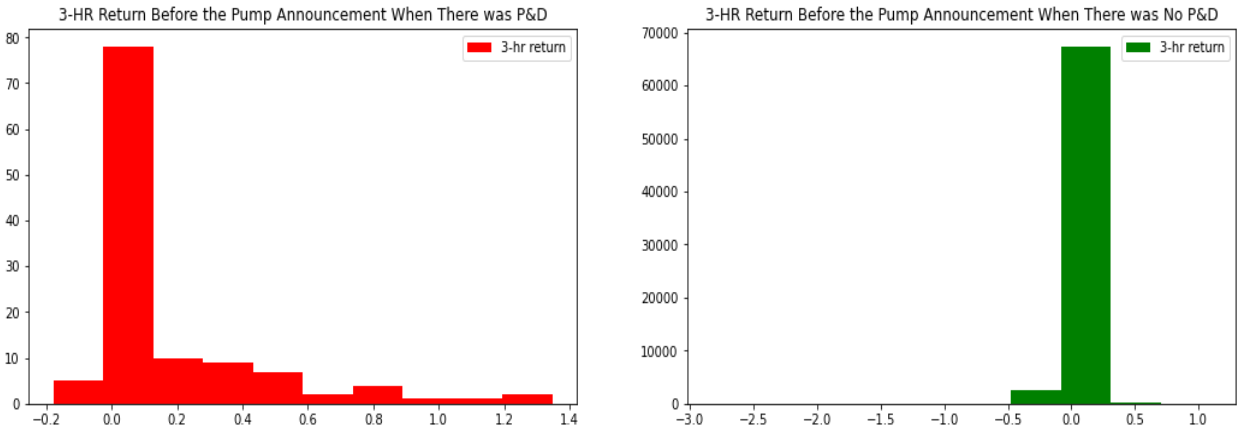


Figure 3. Side by side plot of the 3-hour return before the pump announcement when there was a pump-and-dump scheme and not

Figure 4 represents a side by side plot of the market capitalization with pumped coin's market capitalization on the left side and non-pumped coin's market cap on the right. We can understand from Figure 4 that usually the market capitalization of target pump coins is less than non-pump coins.

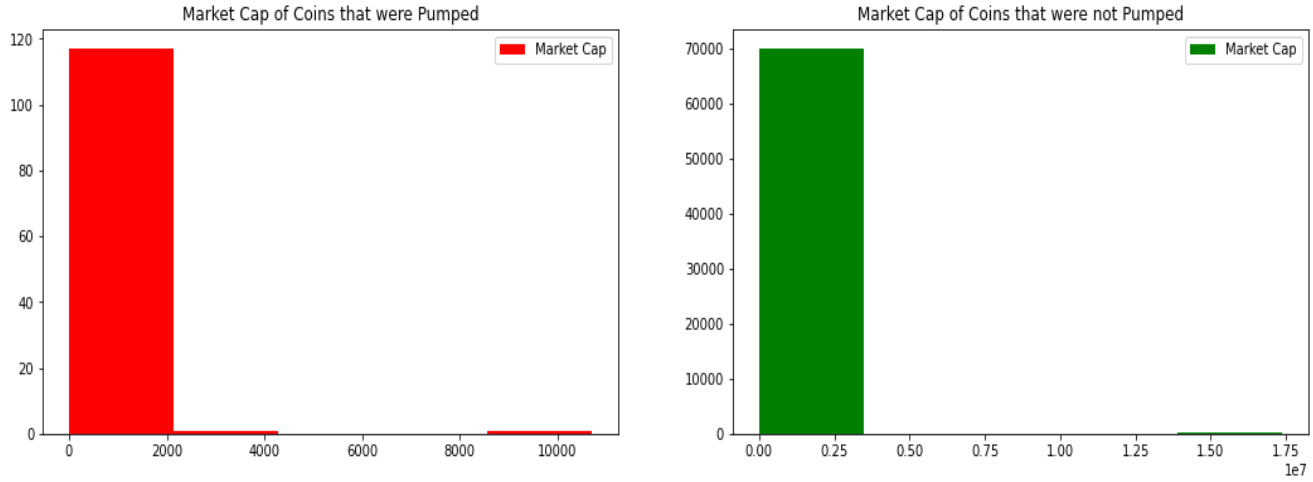


Figure 4. Side by side plot of the market capitalization when there was pump-and-dump scheme and not

Figure 5 represents a side by side plot of the past status of pump with pumped coin's past status on the left side and non-pumped coin's past status on the right. From Figure 5, we can derive that the past status of a pump does not necessarily make the coin subject to another pump in the future.

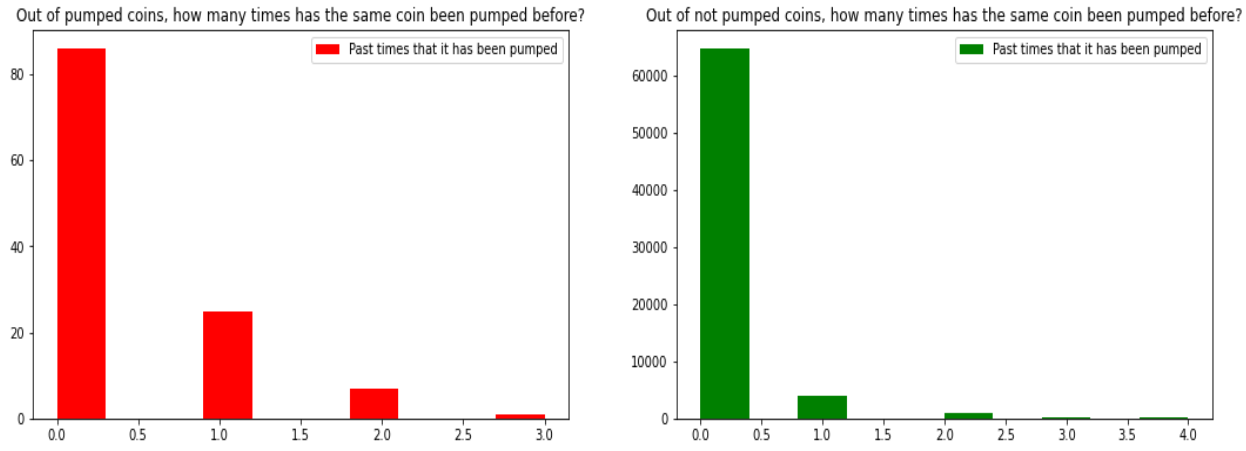


Figure 5. Side by side plot of past pump status of pumped vs non-pump coins

4. Methodology

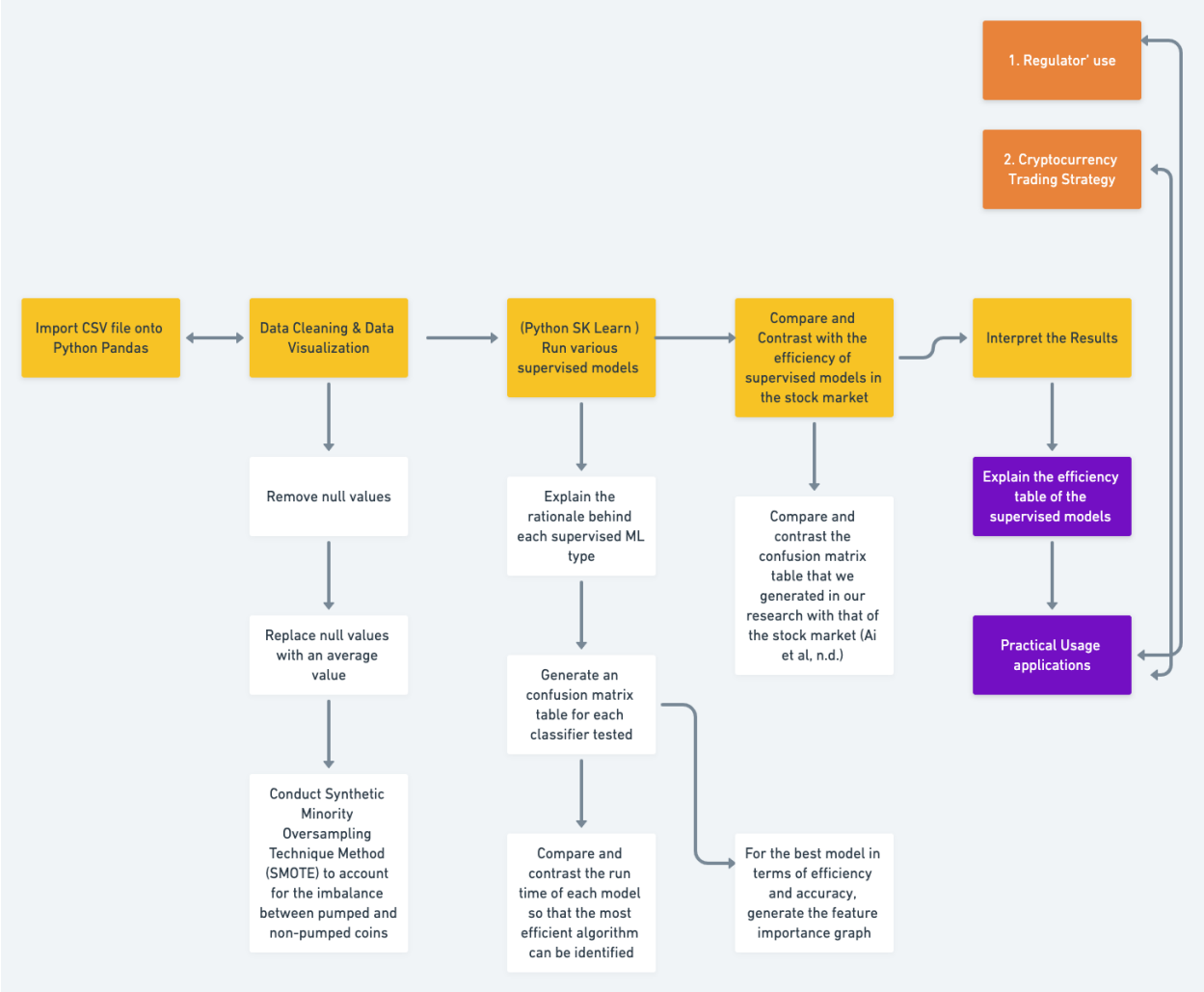


Figure 6. Step by step process of methodology

4.1 Data Cleaning

The first step in data cleaning is removing and replacing null values in the dataset. After this first step and scaling the data points to efficiently run the machine learning methods, one of the biggest problems in the data is the imbalance between the pumped coins and non-pump coins. There are about 119 cases that were pumped in actuality out of 70239 cases. This can cause problems for machine learning models because there is not enough pumped coin's data that

the models can train with. In order to adjust the dataset for this imbalance, we chose to apply the Synthetic Minority Oversampling Technique (SMOTE).

SMOTE is a technique that artificially increases the minority class (in our case, the pumped coin) by synthesizing examples from the minority class. Initially, a random example from the minority class is drawn and then *k* of the nearest neighbors for that example are discovered. A new synthetic example can then be created by running a linear combination of the two randomly selected neighbors (Brownlee 2020).

4.2 Supervised Machine Learning Model Selection

In our research, we want to be able to predict or classify the future pump status of the coin given our technical indicators of a coin. As there are many supervised models available, the efficiency and the accuracy of each supervised model should be tested first. Each classifier model used in our model selection is briefly explained as follows.

4.2.1 Logistic Regression

Logistic regression takes in continuous, independent variables and outputs the binary variable of 0 or 1, which is the probability of a dependent variable occurring. Logistic regression is derived from the sigmoid function (Saini 2021).

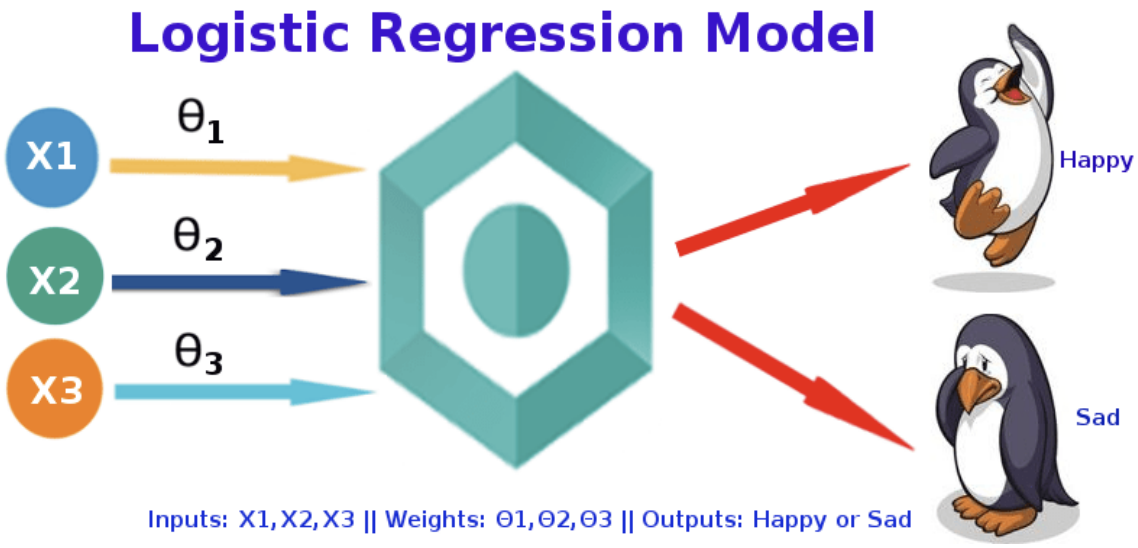


Figure 7. Logistic Regression Description (Saini 2021)

Figure 7 is a simple description of how logistic regression works. Figure 7 shows that logistic regression outputs a binary variable, happy or sad based on certain weights of the input variables.

4.2.2 Gradient Boost Classifier

Gradient boost classifier is an ensemble method that models sequentially and the subsequent models try to reduce the errors of the previous model. Boosting is an ensemble modeling method that is used often for binary classification problems. Boosting algorithms improve the predictive power of the model by converting a multitude of weak learners to strong learners (Saini 2021).

The process of creating a boosting algorithm is by synthesizing a model on the training dataset. Sequentially, another model is built to rectify the errors present in the first model (Saini 2021). This procedure is continued until and unless the errors are reduced as much as possible.

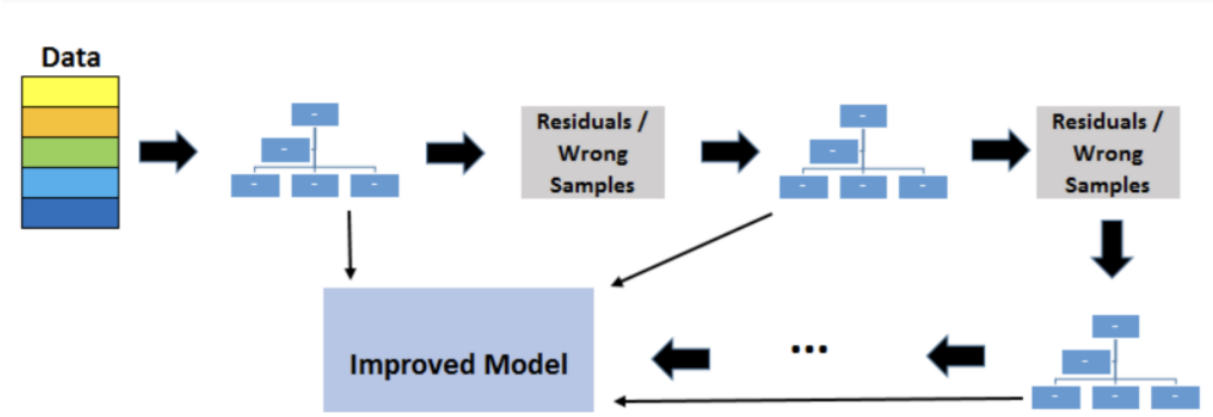


Figure 8. Gradient Boost Classifier Description (Gaurav 2021)

Figure 8 shows gradient boost classifier works by trying to improve upon the residual errors of the previous model.

4.2.3 K-Nearest Neighbor (KNN)

KNN is an algorithm that works by choosing the specified number of K closest to the query. The closest neighbors of the query vote for the most frequent label in the classification

problems (Harrison 2019). Finding the right amount of k to use in the algorithm is also important in the accuracy of algorithms.

4.2.4 Decision Tree Classifier

A decision tree is similar to a flowchart algorithm in that it contains conditional statements that lead to decisions and their probable consequences.

In a decision tree, the leaf nodes correspond to class labels, and the leaf nodes represent the attributes of the class (Goyal 2021).

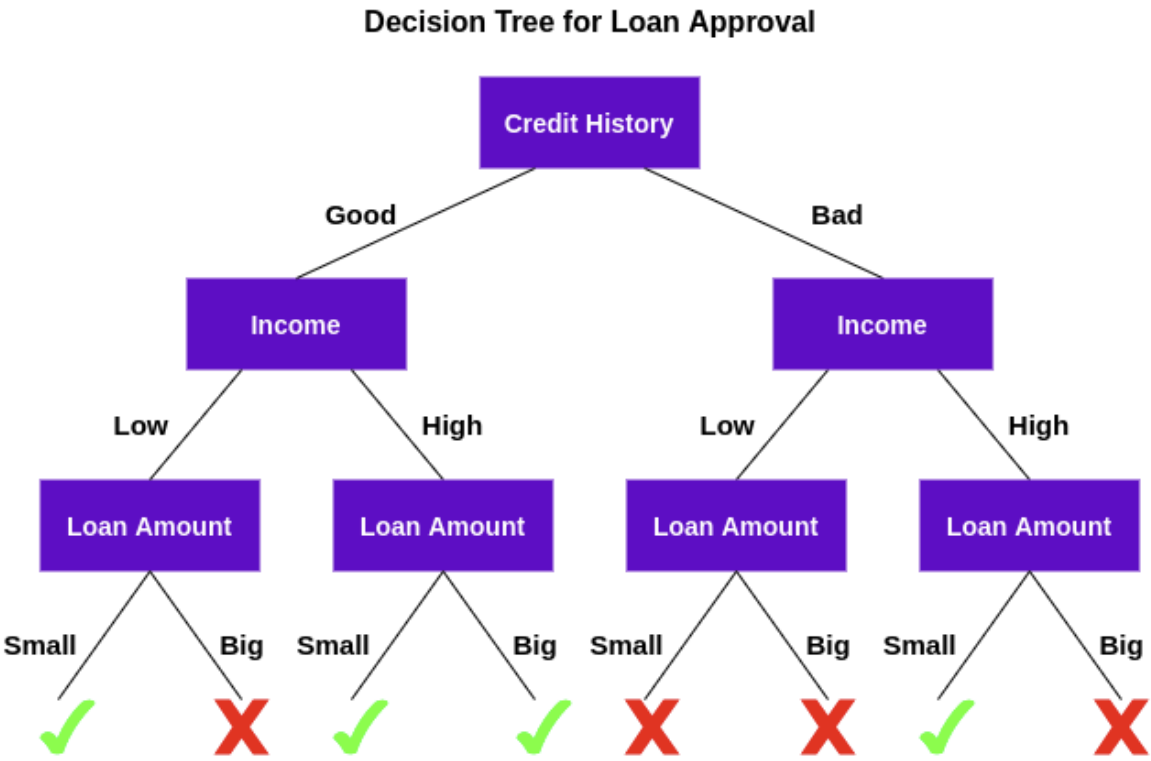


Figure 9. Decision Tree Description (Sharma 2020)

Figure 9 visualizes decision trees simply that trees follow a predetermined order of yes or no questions in order to reach a certain decision in the end.

4.2.5 Random Forest Classifier

Random forests are an ensemble algorithm of many individual decision trees. Random forests reinforce the decision tree's weakness of overfitting likeliness. Also, random forests are called random because each tree in the forest is trained with the bootstrapped data, referring to different samples of training data (Zornoza 2020). Each tree in the random forest is assigned an equal weight in the final output. That is why a random forest is often referred to as the "majority wins" algorithm (Saini 2020).

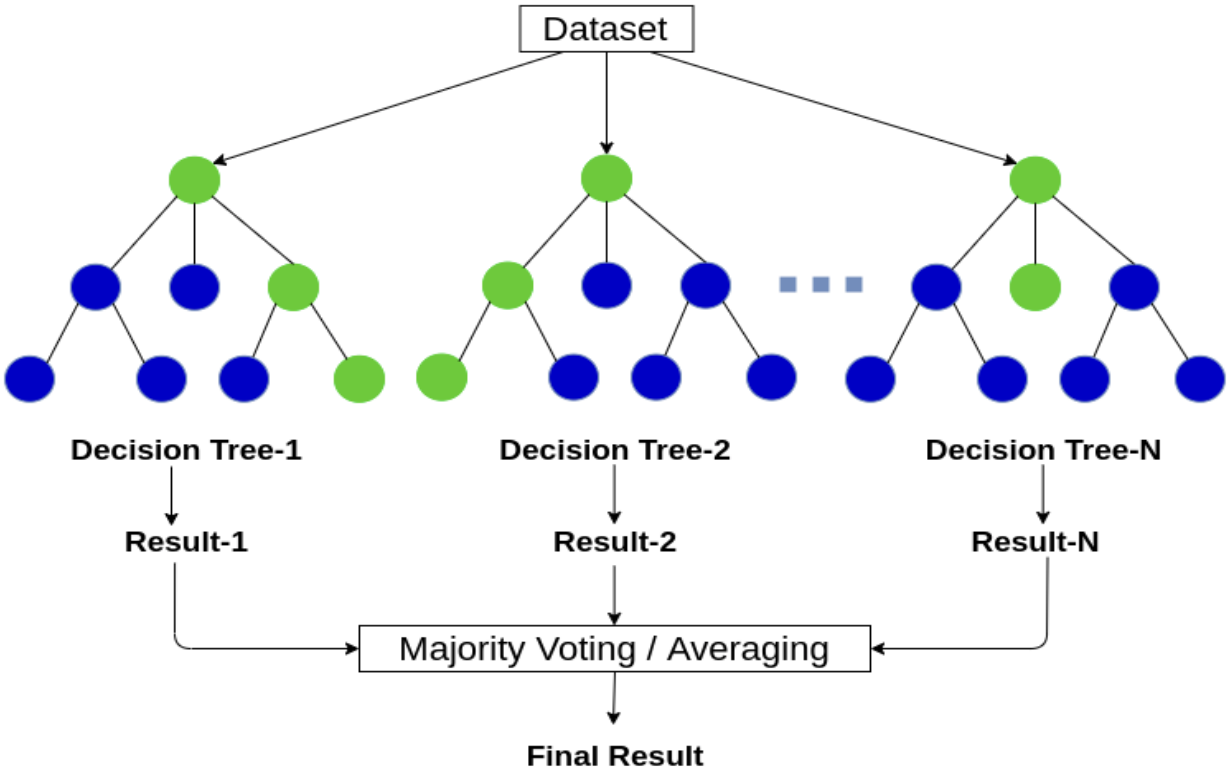


Figure 10. Random Forest Classifier Description (Sharma 2020)

Figure 10 visualizes random forest classifiers to explain that random forest classifiers rely on a multitude of decision trees unlike the decision tree classifier model.

4.2.6 Adaboost Classifier

Like a Gradient Boost classifier, adaptive boost classifier (Adaboost Classifier) is also a type of ensemble classifier that combines several models to improve the ultimate predictive performance of the model. Unlike a random forest and Gradient Boost that assigns equal weights on each tree, Adaboost classifiers assign different weights to each tree. Adaboost classifiers firstly assign *equal weights* to all the data points, but then sequentially adjust weights at the end of creating the next classifier. This is done in such a way that wrongly classified observations have increased weight resulting in their probability of being picked more often in the next classifier's sample (Saini 2020). The resulting, weighted data hence reduces mean squared error as much as possible.

A similarity between Adaboost and Gradient Boost is that they both combine a multitude of weak learners to form a strong learner (Saini 2020). However, whereas Adaboost's wrongly classified observations are fed into the next sequential model by increasing their weights, Gradient Boost tries to fit the next, newly created classifier to the *residual errors* made by the previous classifier. Therefore, the critical difference between Gradient Boost and AdaBoost is what it does with the errors of its predecessors (Choudhury 2020).

4.3 Tools for Analyzing Supervised Machine Learning Models

In "Detecting Market Manipulation in Small–Cap Equities," the authors took a similar method as our research in that they created a dataset after referencing the SEC's confirmed pump and dump cases. (Ai et al. n.d.). The authors inspected over 5,000 civil actions conducted by the SEC between 1996 and 2020. Out of this effort, they found about 8 clear pump-and-dump cases to use in their supervised machine learning model.

The predictors that the authors have used in their supervised models include growth, size, volatility, profitability, dividend yield, and leverage. This is quite different from our research predictors because ours pertain strictly to cryptocurrency technical indicators. Variables such as dividend yield only pertain to the stock market. Regardless of this fundamental difference between the cryptocurrency and the stock market, it is safe to say that the predictors in the model are based on public, readily available information.

Then, Ai and others ran multiple supervised models, including Random Forest, Decision Tree, Support Vector Machine, and Logistic Regression. Ai and others evaluated each model's effectiveness by finding out each model's precision score, recall score, and F1 score.

When a model predicts an output based on the input that is received, the model cannot be 100% correct. For example, when the model is predicting a coin's likelihood of being subject to a pump-and-dump, we can analyze the model's effectiveness with the following evaluation concepts.

- True positives- model correctly predicting a pump-and-dump
- True negative- model correctly predicting non pump-and-dump
- False positive- model incorrectly predicting a pump-and-dump when pump-and-dump does not happen
- False negative-model incorrectly predicting a non pump-and-dump when pump-and-dump happens

With these concepts, we will then be able to evaluate a model's effectiveness. Precision score refers to the percentage of true positives out of the whole predicted positive pool ($\frac{true\ positives}{true\ positives + false\ positives}$) (Shung 2018). Precision refers to the classifier model's ability to find true pumps-and-dumps over a whole number of predicted pumps-and-dumps. Recall score refers to the percentage of true positives out of the whole, actually positive pool ($\frac{true\ positives}{true\ positives + false\ negatives}$) (Shung 2018). Recall score is inherently an ability of a model to find true pumps-and-dumps out of the whole number of true pumps-and-dumps. F1 score is a

harmonic mean between recall and precision ($\frac{recall \times precision}{recall + precision} \times 2$). F1 score ranges from 0 to 1. Conventionally, F1 score of 1 means that the model can classify the dataset perfectly without any false negatives and false positives.

In this research, we aim to use Python's Scikit Learn package, with which we would be able to run multiple supervised machine learning models. When we run supervised machine learning models, we will make use of the k-fold stratified cross-validation to get the most accurate evaluation result on each machine learning's accuracy. As k-fold cross-validation partitions the data in a way that all of our data is at least used once for both training and testing, it is useful in model evaluation (Pramoditha 2020). We will also run the machine learning models without cross-validation so that we can compare and contrast the model evaluation results side by side.

5. Results

5.1 Code Availability

For Python code availability, please refer to the Github link here.

5.2 Supervised Machine Learning Model Evaluation

	Run Time	Average Recall	Average Precision	Average Precision Standard Deviation	Average F1
Decision Tree	0.07	0.964	0.951	0.003	0.971
Random Forest	1.59	0.949	0.965	0.004	0.968
KNeighbors	0.14	0.926	0.912	0.005	0.895
Gradient Boost	2.09	0.821	0.862	0.006	0.884
AdaBoost	0.47	0.759	0.795	0.008	0.837
Logistic Regression	0.11	0.669	0.722	0.008	0.594

Figure 11. Confusion Matrix report of supervised algorithms on the pump-and-dump data (cross -validated with K-stratified fold method (k=10))

Figure 11 is a table that lists each supervised model's average precision, average precision standard deviation, run time, average recall, and average F1 score. Figure 11 represents the result that we were able to achieve with K-stratified fold cross-validation (k=10). Prioritizing a model only based on its F1 score, we were able to derive that Random Forest and Decision

Tree are the only algorithms that achieve a precision rate of over 90%. The algorithm with the lowest F1 score was the logistic regression algorithm with only 59%. The best algorithm in terms of efficiency was Decision Tree classifier because the run time of the Decision Tree was only 0.07 seconds. Despite having a run time of only 0.07 seconds, the Decision Tree classifier's average F1 score was 97%, the highest out of the 6 supervised algorithms that it was evaluated against.

	Recall	Precision	F1
Decision Tree	0.9732	0.9846	0.9789
Random Forest	0.9351	0.989	0.9603
KNeighbors	0.9624	0.8962	0.9281
Gradient Boost	0.8396	0.9373	0.8858
AdaBoost	0.7583	0.9009	0.8235
Logistic Regression	0.626	0.5783	0.6012

Figure 12. Confusion Matrix report of supervised algorithms on the pump-and-dump data (Non-K-fold cross validated score report)

As the authors of "Detecting Market Manipulation in Small – Cap Equities" did not use cross-validation when evaluating each supervised model, we also ran a second round of model evaluation metrics, not using cross-validation. Figure 12 represents the result of non-cross validated model evaluation. As compared with Figure 11, the results do not vary much. The Decision Tree classifier and Random Forest classifier's F1 scores are still the best out of 6 algorithms evaluated. In the non-cross validated version, the Decision Tree's F1 score is 0.0079 higher that of the K-fold cross-validated version of Decision Tree, which was a bit higher than the Random Forest. Similarly, in the non-cross validated version, the Random Forest classifier' s F1 score is 0.968, whereas F1 score is 0.9603 in the cross-validated version of Random Forest.

Therefore, like the Decision Tree classifier, Random Forest's non-cross validated F1 score is a bit higher than that of the K-fold cross-validated version of Random Forest.

However, these differences are still minimal and the final result that the Decision Tree classifier ranks first and then Random Forest ranks second in detecting or predicting pump-and-dump schemes remain unchanged. Also, although Decision Tree has the highest F1 score, as the difference in F1 score between Decision Tree and Random Forest is only within 1 point, we can conclude that both Decision Tree and Random Forest are equally good algorithms to use in predicting pump-and-dumps.

	Recall	Precision	F1
Decision Tree	0.8148	0.8498	0.8319
Random Forest	0.8107	0.9163	0.8603
SVM	0.1523	0.6981	0.2500
Logistic Regression	0.2469	0.7059	0.3659

Figure 13. Confusion matrix report of the pump-and-dump stock market data (Ai et al. n.d.)

Figure 13 represents a confusion matrix report taken from "Detecting Market Manipulation in Small – Cap Equities." From Figure 13, we can derive that the Random Forest classifier and Decision Tree classifier were the best models in terms of F1 score. Unlike our research result (see Figure 11 and Figure 12), the Random Forest model has the highest F1 score and the Decision Tree comes second. Still, Random Forest and Decision Tree's difference in F1 score is only minimal with a 3-point gap.

Having observed that both Decision Tree and Random Forest classifiers are effective in predicting and preventing pump-and-dump schemes in both cryptocurrency and the stock market, we encourage more frequent use of supervised machine learning models in preventing pump-and-dump schemes.

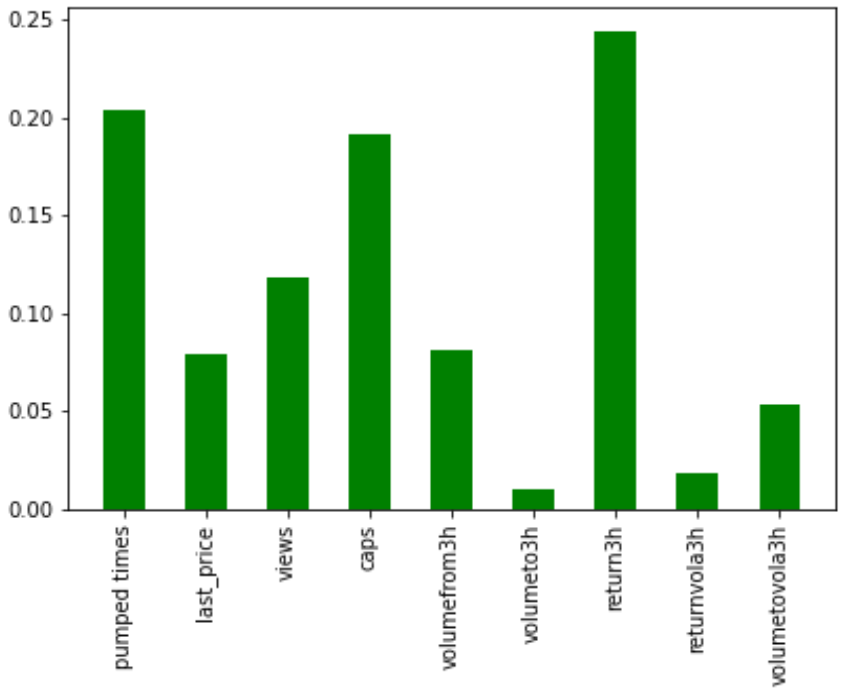


Figure 14. Feature importance graph of the predictors used in the Decision Tree model

Figure 14 is a feature importance graph that shows the best predictor of the pump-and-dump status. From Figure 14, we can derive that the best three predictors of pump-and-dump are the 3-hour price return before the pump announcement, how many times the target coin has been subject to the same scheme before, and the market capitalization of the target coin.

5.3 Result Summary

The bullet points below are the research questions that we raised in section 1.2.

- **What are the features or predictors important in detecting pump and dump schemes in the cryptocurrency market?**
- **What is the best machine learning model for the purpose of predicting pump and dump schemes in the cryptocurrency market?**
- **How effective is the selected supervised machine learning model in detecting pump and dump schemes in the cryptocurrency market?**

- **How does the efficiency of supervised machine learning models compare and contrast in the cryptocurrency market and the stock market?**

In this research, we have found that the best three predictors of pump-and-dump are 3-hour price return before the pump announcement, how many times the target coin has been subject to the same scheme before, and the market capitalization of the target coin. In the cryptocurrency market, we have found that the Decision Tree classifier has the best accuracy with 0.978 F1 score and also is the most efficient with a run time of 0.07 seconds. Despite this result, the Random Forest classifier also has a significant F1 score of 0.96, trailing behind the Decision Tree with a 1-point gap. In the stock market, the Random Forest classifier and Decision Tree classifier were also the best models in terms of F1 score (Ai et al. n.d.). We can safely conclude that supervised machine learning models are useful in detecting pumps-and-dumps in both cryptocurrency and the stock market.

6. Potential Impacts

6.1 Intellectual Merits

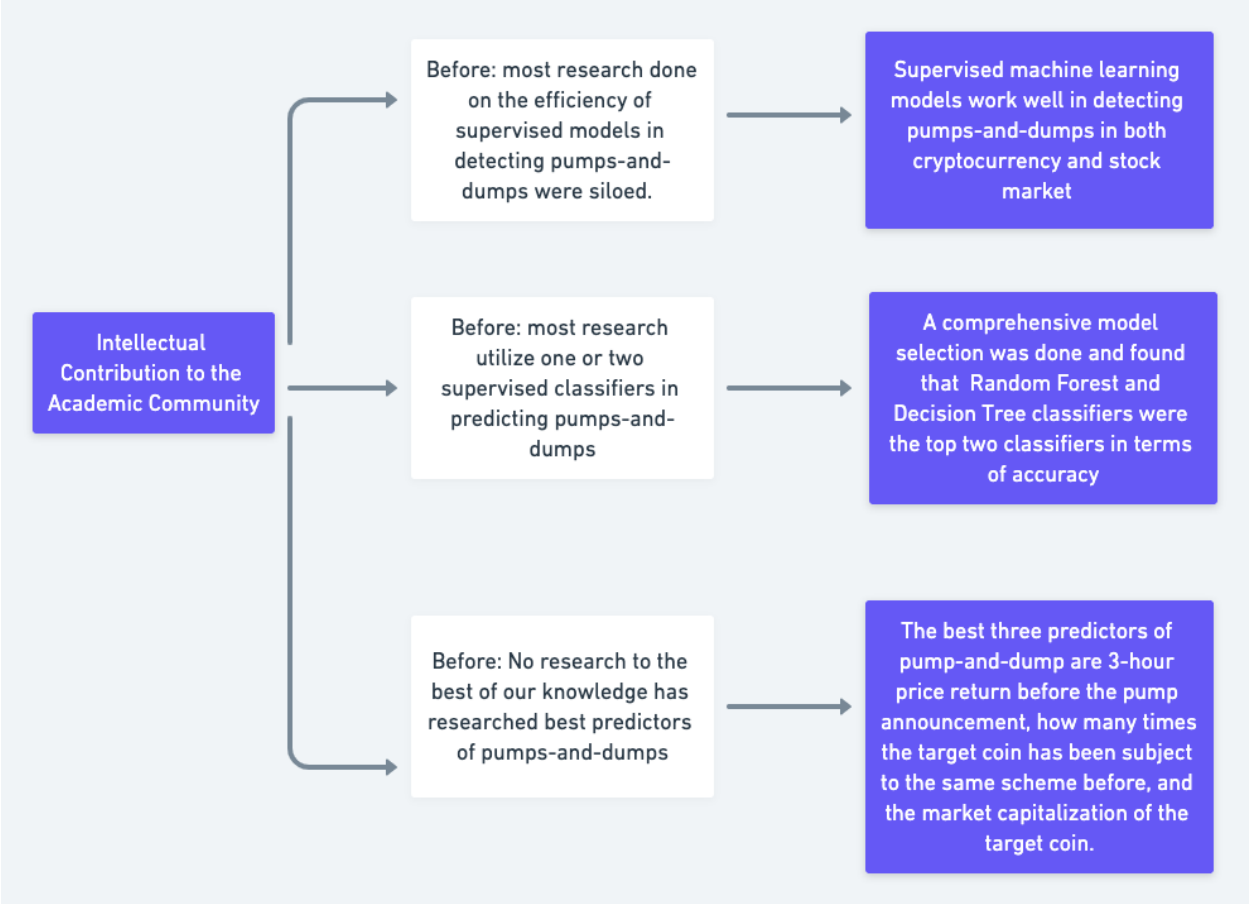


Figure 11. Intellectual Contribution to the Academic Community Flowchart

Figure 11 explains the intellectual insights our research adds to the academic community. From the literature review, we saw that most of the research done on the efficiency of supervised machine learning models in predicting pumps-and-dumps were siloed, meaning that research focused on a particular subset of an asset market. There was research, each focusing on the usefulness of supervised models in cryptocurrency and stock market, but there was a lack of comprehensive comparison on the usefulness of supervised models in both cryptocurrency and stock market. Our research addressed this gap and came to an insight that supervised models can value-add in both cryptocurrency and stock market when predicting pumps-and-dumps.

Furthermore, most literature review research focuses on one or two classifier models rather than conducting a comprehensive model selection to pick the best model for predicting pumps-and-dumps. Our research did a model selection with 6 classifiers to see each model's accuracy and efficiency. Also, our research derived that some of the best predictors of a coin's future pump status are 3-hour price return before the pump announcement, how many times the target coin has been subject to the same scheme before, and the market capitalization of the target coin. To the best of our knowledge, finding the best predictors of pump-and-dump has not been researched as yet before our research.

Despite our research having derived a few meaningful insights, it is not without its limitations. The first limitation of our research is the data size. The number of true pump-and-dump data in our sample is very minimal, which has led to an imbalance in our data, where true pump-and-dump data is small compared to the non-pump-and-dump cases. Although we applied the SMOTE technique to account for this imbalance, our research could have still benefited from a larger data sample.

Another weakness of our research is that we did not actually conduct our own comparative model evaluation of the stock market pump-and-dump schemes. The supervised machine learning model evaluation (see Figure 9) in the stock market is referenced by another author (see Ai et al. n.d.). Although it is unlikely that the final result that Random Forest and Decision Tree classifiers are the best algorithms in detecting pumps-and-dumps would change, it would still be a value-add strategy to conduct a model evaluation on the stock market data that we query.

As our research has demonstrated the effectiveness of supervised machine learning models in predicting a pump-and-dump, our research can potentially inspire further research in the areas of cryptocurrency portfolio management and neural networks model.

Although pump-and-dump is outlawed in the stock market, it is not yet regulated in the cryptocurrency market ("Should Cryptocurrency 'Pump-And-Dump' Schemes Be Regulated?" n.d.). Thus, it means that cryptocurrency investors and traders can still take advantage of either the Random Forest or Decision Tree classifiers in aiding their investment decisions. Pumps-and-dumps organizers tend to lure in uninformed investors with the intention of leaving them in heavy losses after the dump. However, because the target cryptocurrency coin will be undoubtedly pumped, it could potentially lead to good returns. Therefore, further research in portfolio management using pumps-and-dumps would be interesting and would be of value to both academics and cryptocurrency investors.

Another potential area where our research can inspire further research is neural networks. If other academics want to find out the effectiveness of neural networks on predicting or detecting pumps-and-dumps, they could easily apply neural networks to our methodology and derive a meaningful result. As neutral networks are a growing sub-field within machine learning, research on detecting pumps-and-dumps using neural networks would supplement our research.

6.2 Practical Impacts

Our research result has several application scenarios to solve real-world issues such as alleviating uninformed investors' vulnerability to pump-and-dump schemes and improving regulatory engineering. As mentioned in 6.1 Intellectual Impacts, because pumps-and-dumps are not outlawed yet in the cryptocurrency market, it means that cryptocurrency investors and traders can still take advantage of either Random Forest or Decision Tree classifiers in aiding their investment decisions.

Before utilizing supervised models in investment decision-making, investors would have been blind to the presence of pumps-and-dumps schemes. With utilizing supervised models in their investment decision-making, investors can either avoid cryptocurrency coins that are predicted to be targets or make use of the fact that pumps-and-dumps are predicted to occur, giving them a chance to maximize their returns.

So far, pumps-and-dumps are not yet regulated in the cryptocurrency market, but we cannot help but not ignore that pumps-and-dumps have a potential to be regulated in the future ("Should Cryptocurrency 'Pump-And-Dump' Schemes Be Regulated?" n.d.). Our supervised model is quite simple in that it only requires 6 predictors that can be easily queried from cryptocurrency exchanges or information platforms. Therefore, when cryptocurrency pumps-and-dumps become regulated, our model has the potential to help regulators detect pumps-and-dumps schemes and penalize the pumps-and-dump organizers.

Pumps-and-dumps can seriously damage cryptocurrency investors, leaving them with unintended economic losses. The cryptocurrency market should be as transparent as possible and the playing fields should be leveled so that aggressive exploiters cannot prey on the backs of uninformed investors on purpose. As our research has demonstrated the effectiveness of supervised machine learning models in predicting pumps-and-dumps, more practical applications of incorporating supervised models in detecting pumps-and-dumps in the cryptocurrency market would become prevalent in the near future.

References

https://www.mybib.com/j/ObservantFeignedWoodcock

Part II: Supplementary Resources

1. Experts for Comments

- 1. Prof. Samuel W. Buell, buell@law.duke.edu, Duke Scholar
    - Prof. Buell is a professor at Duke Law School and his main research interests are centered around white collar financial crimes. Although this research is focused on using quantitative data to predict the future occurrence of financial crimes, it would still be of great benefit to ask Prof. Buell for comments regarding the possibility of different governments agencies beginning to regulate pumps-and-dumps in the cryptocurrency market.
- 2. Prod. Nuri Bora Keskin, bora.keskin@duke.edu, Duke Scholar
    - Prof. Keskin is a professor at Fuqua Business School and has research interests in financial markets and machine learning. Since he has a lot of research done about applications of machine learning in the financial markets, I think it is going to be of great help for me to get expert comments from Prof. Keskin.

2. Resources for Further studies

- 1. Cryptocurrency and Blockchain (Coursera Open Access Course)
    - This Coursera specialization is a high-level overview of blockchain and cryptocurrency. As my further research areas are cryptocurrency portfolio management that exploits pumps-and-dumps opportunities, it would be useful to listen to this course as this course teaches about the actual process of investing in cryptocurrency and legal regulations around cryptocurrency as of yet.

        - Even for people without much background on cryptocurrency and blockchain, this course provides a good starting point for them to do research in this area.
- 2. Investment and Portfolio Management (Coursera Open Access Course)
    - This Coursera specialization is a high-level overview of portfolio management, using traditional securities such as stocks and fixed income instruments.
    - Although this course does not lecture portfolio management using cryptocurrency, as the basics of portfolio management remains the same regardless of underlying securities, the course is still helpful for further research on cryptocurrency portfolio management.
    - As this course is also geared towards beginners, like the Cryptocurrency and Blockchain course, it provides a good starting point for people to research portfolio management.

3. Seminar, Symposium, and Conference

- 1. USENIX Security' 22
    - Academic conference that is to take place in the August of 2022
    - This conference presents the latest advances in computer systems and network security. As one of my literature review authors has presented a topic on the detection of pumps-and-dumps, this conference would further benefit my knowledge of computer security and its relevance with cryptocurrency pumps-and-dumps.
- 2. Compliance & Financial Crime Conference
    - Academic conference that is geared towards financial compliance officers, risk officers, and computer security officers.
    - This conference focuses on many forms of imminent and real compliance threats that financial firms face today. As fraud detection is one of the main focus areas in this conference, attending this conference would broaden my knowledge on how fraud detection is processed within financial companies. Potentially, this research

could also contribute to this conference as well since there is a significant relevance in the subject area.

Part III: Related Products

1. Experiential Learning Activities (Link to the Full EL Report Form)

- 1. Online conference with Prof. Jimmy Lenz from Duke on decentralized finance (DEFI) and blockchain
    - This online conference with Prof. Lenz contributed to my knowledge of DEFI and how DEFI is disrupting the traditional finance sector. Because Prof. Jimmie Lenz, himself, has a number of years that he spent in the traditional banking sector, he was able to share his experience working in both DEFI and centralized finance (CEFI). His guest lecture also inspired me to think about the long-term impacts of DEFI and the future of finance.
    - As DEFI becomes more commercialized and more people get to learn about its advantages of DEFI (transparency, access, equity, high efficiency, and etc..) might even surpass the market share that centralized finance now holds in the finance industry.
    - Before Prof. Lenz's lecture, I was quite new to the topic of DEFI and how fast the cryptocurrency market is growing. This lecture contributed a lot to my subject knowledge and guided me in narrowing down my SW research focus area.
- 2. 7th Blockchain Innovation Summit hosted by Wanxiang Blockchain Lab
    - This online summit featured Vitalik Buterin, founder of Ethereum protocol where he talks about the major upgrades in Ethereum 2.0 and the future of the network.
    - This conference, in general, was very helpful for me to get the feel of the whole industry in the blockchain. I could sense that most industry leaders and professionals believe in the enormous potential of blockchain.

- The 7th Blockchain Innovation Summit hosted by Wanxiang Blockchain Lab provided me with a good opportunity to listen in to a talk by the founder of one of the biggest and arguably, one of the most important blockchain protocols, Ethereum.
- Like the online conference by Prof. Lenz, this online summit guided me in narrowing down my SW research focus area to pumps-and-dumps in the cryptocurrency market.

2. Seminar, Symposium and Conference Presentations

- Class of 2022 SW Conference
    - Date: April 23rd, 2022
- As an extended product of my signature work research, this research will be displayed at the Duke Kunshan Class of 2022 SW Conference through the means of an online presentation video.
- Powerpoint Presentation Link

3. Publications

- None available

4. Fellowship, Grants, Offers

- None available

Part IV: Signature Work Documents

[Include URLs to all the signature work documents submitted to the signature work office. If the document is not applicable, leave the information blank.]

Document type	URL to PDF	date submitted
SW Declaration of Intent	In the google folder	
SW Mentor Agreement Form	In the google folder	
Team-Based Project Agreement Form		
SW Project Proposal Form	In the google folder	
SW Experiential Learning Proposal Form		
SW Experiential Learning Report Form	In the google folder	
SW Experiential Learning Supervisor Report Form	In the google folder	
RCR Certificate	In the google folder	
Petition to Change SW Project Proposal Form		
Capstone Course Schedule Change Request		